DOCUMENT RESUME

ED 333 045 TM 015 614

AUTHOR Cook, Nancy R.; And Others
TITLE Evaluation and Validation: A Look at the Program Effectiveness Panel.
PUB DATE 91
NOTE 14p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Data Analysis; Educational Assessment; Elementary Secondary Education; *Evaluation Methods; *Evaluators; Federal Government; Government Role; *Program Effectiveness; Program Evaluation; *Research Design; *Validated Programs
IDENTIFIERS Department of Education; *Program Effectiveness Panel; Validation Verification and Testing Techniques

ABSTRACT

The work of the Program Effectiveness Panel (PEP) of the United States Department of Education was examined. The PEP is the Department of Education's primary means of certifying or validating the effectiveness of educational programs. Thirty-three educational programs were validated by the PEP between January 1988 and December 1990. Over time, about 60% of applicant programs have been approved. A content analysis of successful program submittals and summaries of panelists' comments indicated that claims of student achievement were frequent and that the domains of achievement varied widely from basic academic areas (such as reading, mathematics, and science) to knowledge of such content as parenting practices, the visual arts, the law, and problem solving skills. Sources of evidence varied as well, from commercial tests and performance assessment to observation and interview data, school records, and attitudinal instruments. Several flaws in evaluation design in unsuccessful submittals were found by PEP members. These often included: (1) failure to assess the persistence of effects; (2) small magnitude of results; and (3) incorrect or inadequate analytic procedures. The PEP is notable for its focus on measured effects of programs. Its validation is an important component of federal efforts to disseminate exemplary programs and practice. Attachment A lists the name, claims summary, instruments used, and comparisons for each of 13 projects approved/validated between January 1988 and December 1989 by the PEP. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED330045

Evaluation and validation: A look at the Program Effectiveness Panel

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Nancy R. Cook
M. Christine Dwyer
RMC Research Corporation

Charles Stalford
O.E.R.I./ U.S. Department of Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Nancy R. Cook
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SUMMARY

This paper describes the U.S. Department of Education's Program Effectiveness Panel (PEP) and recent trends in its evaluation of exemplary educational programs. The PEP [1] serves as the U.S. Department of Education's primary means of certifying (or validating) the effectiveness of educational programs. All programs in the National Diffusion Network (NDN) are validated through this process before being nationally disseminated in the network. Specifically, this paper addresses five questions: [1] What are the claims of effectiveness made by educational programs that have been validated in recent years?; [2] What types of evidence are used to support program claims?; [3] What characterizes the evaluation designs of successful submittals?; [4] What are some alternatives to traditional comparison standards used in successful submittals?; and [5] What are the problems identified by panelists when programs have been unsuccessful? The paper also describes the evaluation designs and results of several recently successful projects.

Findings: Thirty-three educational programs were validated by the PEP between January 1988 and December 1990. The percent of applicant programs that pass has averaged approximately 60 percent over time. The data used for the findings consist of successful program submittals and summaries of panelists' comments both for successful and unsuccessful projects. A content analysis of successful submittals and summaries of panelists' comments was conducted for this paper. Claims of student achievement remain very frequent (23 of the projects made a claim in this area); however, the domains of achievement vary widely from basic academic areas, such as reading, mathematics and science, to knowledge of such content as parenting practices, the visual arts, the law, problem-solving skills. Claims based on teacher behaviors and student attitudes appear less frequently but are found in several submittals. Evidence sources vary from commercially available tests, project-developed tests, and performance assessment to observation data, interview data, school records, and attitudinal instruments. Many projects use alternative comparison standards including contrasting the results of program participants to norming samples, national statistics or research-based standards. Flaws found by panelists in the evaluation designs of unsuccessful submittals include: presumed important effects not measured; difficulties in attributing the effects to the intervention; inadequate comparison standards or instrumentation; or reliance on formative information, rather than outcomes. In the area of the results of unsuccessful projects, panelists usually commented on the projects' failure to assess the persistence of effects; small magnitude of the results; or incorrect analytic procedures. In the area of transportability, the context for unsuccessful projects often seemed highly unusual. Unlike many federal programs that examine programmatic features, the PEP is singular in its focus on measured effects. PEP validation remains an important component of federal efforts to disseminate exemplary educational programs and practices.

---

[1] The current form and name were established in 1987. The PEP was preceded by the Joint Dissemination Review Panel (JDRP) which had operated since 1972.

TM016614

BEST COPY AVAILABLE

Evaluation and validation: A look at the Program Effectiveness Panel

Nancy R. Cook
M. Christine Dwyer
RMC Research Corporation

Charles Stalford
O.E.R.I./ U.S. Department of Education

Introduction: This paper describes the U.S. Department of Education's Program Effectiveness Panel (PEP) and recent trends in its evaluation of exemplary educational programs. All programs in the National Diffusion Network (NDN) are validated through this process before being nationally disseminated in the network. Specifically, this paper addresses five questions: [1] What are the claims of effectiveness made by educational programs that have been validated in recent years?; [2] What types of evidence are used to support program claims?; [3] What characterizes the evaluation designs of successful submittals?; [4] What are some alternatives to traditional comparison standards used in successful submittals?; and [5] What are the problems identified by panelists when programs have been unsuccessful?

The Program Effectiveness Panel (PEP) serves as the U.S. Department of Education's primary means of certifying (or validating) the effectiveness of educational programs. Its current form was established in 1987. [1] The PEP is composed of approximately sixty individuals with expertise in program evaluation. One-third of its membership are federal employees while the remainder are associated with institutions of higher education, school districts, professional associations and other entities concerned with educational research, evaluation and practice. The PEP is integrally connected to the National Diffusion Network. Since all projects in the NDN have been validated, individuals in the network are promoting "proven" products, rather than simply programs with innovative features. This difference is important to school districts and other educational institutions when considering changes in their programs.

Making the Case (Ralph and Dwyer, 1988) describes the PEP criteria and guidelines for programs. It discusses, in more detail, the background and purposes of the PEP; how the panel works; how standards and assumptions have changed; the criteria of effectiveness; and types of claims programs usually make. It also includes the format for submissions to the panel and a list of evaluation references. The three criteria upon which the PEP judges projects are: [1] evaluation design (up to 40 points); [2] meaningful results (up to 50 points with a minimum of 40); and [3] transportability (up to 10 points). These three criteria combine for a possible total of 100 points, of which projects must accumulate a median score of 70 to pass.

A wider range of educational evaluations have been successfully used for validation since the change in the criteria reflected in Making the Case. Attachment A (Summary of Projects Approved by PEP) lists all projects that have been validated since January 1988. The successful projects described in the next section faced various challenges in evaluation design or in the interpretation of the meaningfulness of results.

Description of selected projects: Jeffco Life Science, a junior high life science project, used a quasi-experimental design with two groups taking pre and posttests. For the evaluation, the

---

[1] The PEP was preceded by the Joint Dissemination Review Panel (JDRP) which had operated since 1972.

project chose "experimental" schools from the schools where the science teachers had used the curriculum for one year. This sample was chosen in order to represent the diversity of socio-economic status within the district. Comparison schools were chosen from the area to match the SES distribution of the schools. Science teachers in the comparison schools were interviewed to ensure that they taught the concepts and facts measured by the project developed science tests. A standardized test of achievement was also administered as an independent measure of student achievement. The statistical analysis revealed treatment effects that were independent of teacher effects and were persistent for all measures for two years. The project argued that its results were meaningful on the basis of the magnitude of the effects and the type of learning (i.e., more in-depth and conceptual) being qualitatively better.

The evaluation design for Sound Foundations was a variation on the experimental/control pre-post design. This ninth grade remedial math curriculum project included three groups in its evaluation study - Group 1 received the complete intervention (including daily math quizzes); Group 2 completed the daily quizzes but received the school's general mathematics curriculum; and Group 3 simply received the general curriculum with no daily quizzes. Intact math classes in participating schools were assigned to one of the three groups. Standardized math tests and attitudinal instruments were given to all students as pre and posttests. One potential argument for the program's effectiveness was defused by this design - daily testing with any curriculum would increase achievement. The submittal presented evidence that the entire program (including the daily quizzes) had greater effects than "ordinary" curricula with daily quizzes and without the quizzes. The argument for meaningfulness of the results rested not only on statistical differences but on the greater academic need of the target population.

Project ADAPT, a comprehensive service delivery system for secondary and post-secondary learning disabled students, compared descriptive information about participants to national statistics about this population, local statistics about non-LD students, and expectations derived from other research studies about the LD population. Percentages of ADAPT students being mainstreamed were compared to national statistics. A second effect concerning the percentage of ADAPT students dropping out of school included comparisons to national dropout rates of all students and LD students and to the local dropout rate of all students. Finally, on five standardized tests of achievement, longitudinal data were analyzed for participating students for the five years of the program operation. Effect sizes for ADAPT students were compared to independent research studies of LD students. These comparisons, although not internal to the project (e.g., a project comparison group), all argued for the meaningfulness of ADAPT's descriptive results for a group of students with special needs.

GRADS (Graduation, Reality and Dual-Role Skills) compared participating Ohio teens who are parents to national statistics about adolescent parents, including their drop-out rate and the birth weight of their babies. For claims about dropping out of school and the health of the babies of participants, the project also included an analysis of how time spent in the program related to the effects demonstrated (i.e., the longer time in the program, the greater the benefits). Finally, the project included pre and posttest data concerning some participating students' knowledge about parenting skills and child development. The project argued persuasively about the severity of the problems associated with teenage parenthood and the importance of having programs with documented effects, rather than arguing for the magnitude of their effects.

On the Way to Success with Early Prevention of School Failure used a longitudinal study to show that students who had participated in EPSF in kindergarten and then continued in the first grade year demonstrated persistent achievement effects into second grade. Two standardized achievement tests were administered each school year to the participating group and a small

comparison group. The program argued for the meaningfulness of results in terms of magnitude of gains but also in terms of the importance of maintenance of the gains made by at-risk students.

Claims of effectiveness: The PEP judges programs on the basis of the claims that the program makes about its effectiveness. Claims are summaries of the observable effects of a program. They are categorized by type of effects (e.g., cognitive, affective, or behavioral). Making the Case outlines four major claim types and discusses them at length in Chapter 5. Programs may make claims other than cognitive gains (such as changing teachers' practices) but must argue that their effects contribute eventually to student learning.

As can be seen from Attachment A as well as the description of the projects in the previous section, there is a wide range of claim types related to student achievement. Achievement remains the most frequent claim although claims related to behaviors and attitudes are also increasingly represented in the group of successful projects.

Types of evidence: There is a clear trend toward use of project-developed measures. With these measures, projects usually argue that the measure they have developed is more appropriate for the intervention, and the data are therefore more valid indicators of the program's effects. Projects that depend on project developed tests must counter the argument that the comparison students had no opportunity to learn the content included in the test. One of the previously described projects, Jeffco Life Science, was particularly thoughtful about this problem, including concepts that comparison teachers identified as ones that they taught in the four science tests developed by the project.

Evaluation designs: Many projects use quasi-experimental designs to evaluate program effects. Other projects use descriptive studies and compare results to comparison standards that are published or generated by research. Fundamentally, a good evaluation design convinces the panelists that the effects that are included in the program's claims could reasonably result from the program's intervention.

Alternative comparisons: As can be seen from the earlier descriptions of projects, descriptive evaluation designs without traditional comparisons are increasingly accepted by the panel. Nonetheless, projects must demonstrate to the panel that results are somehow meaningful in a comparative sense. The most usual method has been to compare project participants to comparable individuals who have not received the intervention. Meaningfulness then is based on the magnitude of the results. A second type of comparison standard depends on norming samples. If a project uses a norm-referenced standardized test as a measure, then it may compare the gains made by participating students (pre to post) to the gains that are normal for a nationally representative group of similar students. A project that included a norming sample for comparison purposes was the New Jersey Youth Corps. National statistics or research about certain populations of students may also provide projects with comparison standards. As previously mentioned, ADAPT used other individuals' research about LD students as its comparison while GRADS, a program for pregnant and parenting teens, used national statistics about this group as a comparison standard. When projects use alternative comparison standards, they often argue for the meaningfulness of results based on the importance of the need they are addressing. Successful projects with statistically significant results usually argue that their results are meaningful in an educational sense as well. Multiple standards greatly strengthen a program's case since any one comparison standard may be questioned by the panel.

Why do some PEP candidates fail?: As was mentioned earlier, panelists judge projects on the basis of three criteria: evaluation design, meaningful results, and transportability. In the area of design, there are five often cited flaws. These are: [1] that presumed important effects of the program are not measured by the design; [2] that there was no reasonable way to directly attribute the effects observed to the program and other causes of change are highly plausible; [3] that the comparison standard is inappropriate or the comparison sample is small in size, non-representative, and/or "do-nothing" in the sense of no fair opportunity to learn/develop; [4] that the instrumentation is weak or unrelated to the outcomes and programs; and [5] that the evaluation is formative in nature rather than providing evidence of program outcomes. In the area of meaningful results, panelists often comment on: [1] lack of attention to the persistence of effects; [2] results that are minor in light of the scope of the program; [3] results that are not significant in the majority of cases; or [4] analytic procedures that are incorrect. In the area of transportability, the most frequent comment is that the context for unsuccessful projects seems highly unusual.

Conclusions: In the period since the PEP changed its criteria, program evaluations validated by the panel have become more varied. The approaches to evaluating the impact of functioning real-world programs have similarly become more valid. The PEP process has adapted to the variety while maintaining a high standard for judging programs on the basis of outcomes. Unlike many federal programs that examine programmatic features, the PEP is singular in its focus on measured effects. PEP validation remains an important component of federal efforts to disseminate exemplary educational programs and practices.

For more information: Readers interested in learning more about the National Diffusion Network may contact:

National Diffusion Network
U.S. Department of Education
Recognition Division
O.E.R.I./P.I.P.
555 New Jersey Ave, NW
Washington, DC 20208
202-219-2134

For information about the PEP and its validation process, contact:

NDN Technical Assistance Center
RMC Research Corporation
400 Lafayette Rd.
Hampton, NH 03842
800-258-0802

ATTACHMENT A

Summary of Projects Approved by PEP

SUMMARY OF PROJECTS APPROVED BY PEP

January 1990 - December 1990

This is a very abbreviated description of the evaluation characteristics of the thirteen projects approved last year. Asterisk indicates project or project variation has been approved previously.

PROJECT NAME	CLAIMS SUMMARY	INSTRUMENTS	COMPARISONS
1. SSBD	■ screening procedure is reliable predictor; screening procedure accurately discriminates behavior disordered from students not at risk	■ various approaches to analyzing basic SSBD, including multiple observers; internal scale consistency	■ interobserver reliability, stability over time, interscale comparisons
2. Jeffco Life Sciences	■ changes in life science concepts, knowledge, applications	■ four science tests	■ comparison group
3. Sound Foundations	■ gains in basic math skills	■ math achievement test	■ comparison with competing treatments
	■ positive change in attitude toward success in math; females equally positive	■ rating scale	■ comparison with competing treatments, subgroup comparison
	■ change in subsequent math course selections	■ 9 years of course records	■ historical trend
	■ increase in pass rates	■ state competency test	■ historical trend
4. ADAPT*	■ increase in proportion of LD students mainstreamed	■ student records	■ national estimates
	■ decrease in dropout rate of LD students	■ school records	■ national estimates
	■ gains in basic academic skills	■ composites of variety of tests	■ before-after, effect size

PROJECT NAME	CLAIMS SUMMARY	INSTRUMENTS	COMPARISONS
5. GRADS	■ increase in pregnant teens retained in school	■ school records	■ national averages
	■ gains in knowledge of positive parenting practices	■ project-developed test	■ before-after-3rd point
	■ higher birth weights (healthier babies)	■ hospital records	■ state averages and length of participation comparisons
6. Books and Beyond*	■ decrease in TV viewing, increased discrimination in viewing	■ parent and student surveys, logs	■ pre-post, compared to original comparison group
	■ increase in choice of reading over TV, increase in reading time and number of books read		■ trend of results compared to national estimates
7. New Jersey Youth Corps	■ gains in basic skills	■ TABE	■ norming sample
	■ job placement success of graduates, diploma attainment, hours of participation	■ records	■ state and national results from competing treatments
8. KITE*	■ gains in basic skills	■ CTBS	■ norming sample, comparison group
9. First Level Language	■ gains in language concept development	■ CPI, Boehm, CSAB	■ norming sample, comparison group, trend of passing level over time
10. ECRI*	■ gains in reading skills for different student populations	■ variety of standardized achievement tests	■ norming sample, comparison group
11. Study Skills Across the Curriculum*	■ gains in study skills	■ project-developed test	■ comparison group
	■ gains in behaviors	■ project-developed self-report inventory	■ comparison group
	■ improved performance in content courses	■ unit tests and grades	■ comparison group

PROJECT NAME	CLAIMS SUMMARY	INSTRUMENTS	COMPARISONS
12. First-Level Math*	■ gains in math achievement	■ MRT, CIRCUS, CTBS	■ norming sample ■ disaggregation by program variation
13. On the Way to Success in Reading and Writing*	■ gains in language and basic skills ■ reductions in retention, special ed. referrals	■ PPVT, ITBS ■ school records	■ norming sample, comparison group ■ historical trends

SUMMARY OF PROJECTS APPROVED BY PEP

January 1988 - December 1989

This is a very abbreviated description of the evaluation characteristics of the twenty projects approved in the last two years. Asterisk indicates project or project variation has been approved previously.

	PROJECT NAME	CLAIMS SUMMARY	INSTRUMENTS	COMPARISONS
1.	FOR SEA*	■ gains in content knowledge of marine science	■ project developed tests	■ comparison group
2.	Pierce County	■ enhancement of rate and quality of employment and post-secondary enrollment	■ school records ■ structured interviews	■ historical pattern ■ comparison group
3.	CIRC	■ gains in reading achievement	■ norm-referenced achievement tests, informal reading inventory	■ multiple comparison groups
4.	Effective Citizenship	■ gains in knowledge of law ■ increases in favorable attitudes toward school, police, law-abiding behavior ■ reductions in delinquent behavior	■ project-developed questionnaires and tests of knowledge	■ control groups with competing treatments
5.	HOTS	■ gains in reading and math basic skills ■ gains in higher order thinking skills ■ changes in self-concept	■ achievement tests, locally-developed criterion-referenced test ■ Ross Test, Inference from Context Test ■ anecdotal reports from teachers	■ norming samples, comparison group ■ comparison group
6.	Visual Arts	■ changes in teachers' instructional strategies ■ increase in students' skills ■ institutionalization of practices	■ structured interviews, naturalistic observations ■ project-developed tests ■ records of local diffusion	■ comparison teachers ■ comparison group, high-low implementor comparison

PROJECT NAME	CLAIMS SUMMARY	INSTRUMENTS	COMPARISONS
7. Mechanical Universe	■ gains in physics achievement ■ increases in positive attitudes toward physics ■ positive teacher attitudes about topic exposure and need for curriculum	■ project-developed tests ■ teacher and student questionnaires	■ control groups, quartile group comparisons
8. Sharing Successful Programs	■ cost-effectiveness of model for transferring effective programs	■ records of participation and maintenance, interviews, financial estimates	■ historical comparisons, comparative cost analyses
9. Know Your Body	■ gains in health knowledge ■ changes in health behaviors ■ changes in physical risk factors (e.g., fat intake, reduction, endurance, lowered cholesterol)	■ tests ■ questionnaires ■ physical measurements	■ control groups
10. INSITE	■ gains in cognitive and motor development, socialization, daily living ■ enhanced rate of home placement rather than institutionalization ■ change in parent skills and behaviors	■ Callier-Azusa test ■ records ■ survey, records	■ developmental scale norms and historical patterns ■ data from national studies
11. KIDS KITS for Middle School*	■ improvements in research-related skills and behaviors (e.g., posing questions, use of variety of resources, etc.) ■ increase in interest for research activities	■ structured interviews	■ high-low implementor comparison, previous comparison group results

	PROJECT NAME	CLAIMS SUMMARY	INSTRUMENTS	COMPARISONS
12.	Child Development	■ enhanced prosocial attitudes and behaviors (e.g., consideration of others' needs and feelings)	■ observations, teacher and student questionnaires, student interviews	■ control classrooms
13.	Success Enrichment*	■ enhanced writing skills	■ writing samples	■ control groups
14.	COMP	■ increased use of effective management practices by teachers	■ structured observations	■ comparison groups with competing treatments
		■ reductions in student off-task behavior	■ questionnaires	
		■ gains in English and math achievement	■ achievement tests	
15.	Videodisc Instruction in Math	■ gains in math achievement	■ achievement tests	■ controls with competing treatments, norming sample
		■ change in teacher skills with special education students	■ interviews, observations	■ regular and special education results
16.	CERES*	■ gains related to career awareness and development	■ career education tests and inventories	■ comparison group
		■ enhanced skill development	■ Adult Performance Level Survey	
		■ enhanced employability	■ employment records	
17.	DeLaSalle	■ improved attendance	■ school records	■ historical comparisons, districtwide averages
		■ increase in academic competencies	■ TABE	■ comparison group
		■ changes in educational attitudes and long-term employment outcomes	■ structured interviews, testimony	■ comparison group

	PROJECT NAME	CLAIMS SUMMARY	INSTRUMENTS	COMPARISONS
18.	Learning to Teach in Inner-City Schools	■ increase in frequency of positive teaching behaviors among student teachers and supervising teachers	■ structured observations	■ comparison groups of student teachers
		■ reduction in "off-task" behaviors among students of student teachers		
		■ participants are more likely to remain teaching in an inner-city environment	■ telephone interviews	
		■ participants receive more positive ratings from principals		
19.	Social Decision-Making	■ improvements in problem-solving and social skills	■ criterion-referenced tests	■ comparison groups, high-low implementation comparisons
		■ increases in pro-social behaviors and abilities to cope with stress	■ semi-structured interviews, teacher ratings, sociograms	
		■ long-term reductions in negative behavior	■ National Youth Survey, Youth Self Report	
		■ increase in teachers' facilitation skills	■ observations	■ comparison group teachers
20.	FIST*	■ tutors of adults increase teaching skills	■ project-developed test	
		■ increase in adult reading achievement	■ achievement tests	■ norming sample
		■ positive changes in adult life experiences	■ survey	